Developing & Scaling Research Data Management (RDM) & Curation Services

Tim McGeary
Associate University Librarian
Duke University

Sophia Lafferty-Hess
RDM Consultant
Duke University

Claire Stewart
Associate University Librarian
University of Minnesota

Jennifer Darragh
RDM Consultant
Duke University

DUKE UNIVERSITY LIBRARIES

LIBRARIES
UNIVERSITY OF MINNESOTA

The Duke Story

Reasons to Scale Up RDM and Curation

2011 - NSF requires Data Management Plans

2013 - OSTP memo requiring Federal agencies with more than $100M in R&D expenditures to develop plans to make the results of federally funded research freely available to the public—generally within one year of publication.

2016 - NSF and Dept of Energy begin requiring deposit of publications and require data to be made available at expense of research institutions

The History at Duke

2015 - NSF rejects Duke research proposal due to insufficient DMP, specifically due to lack of plans to deposit data

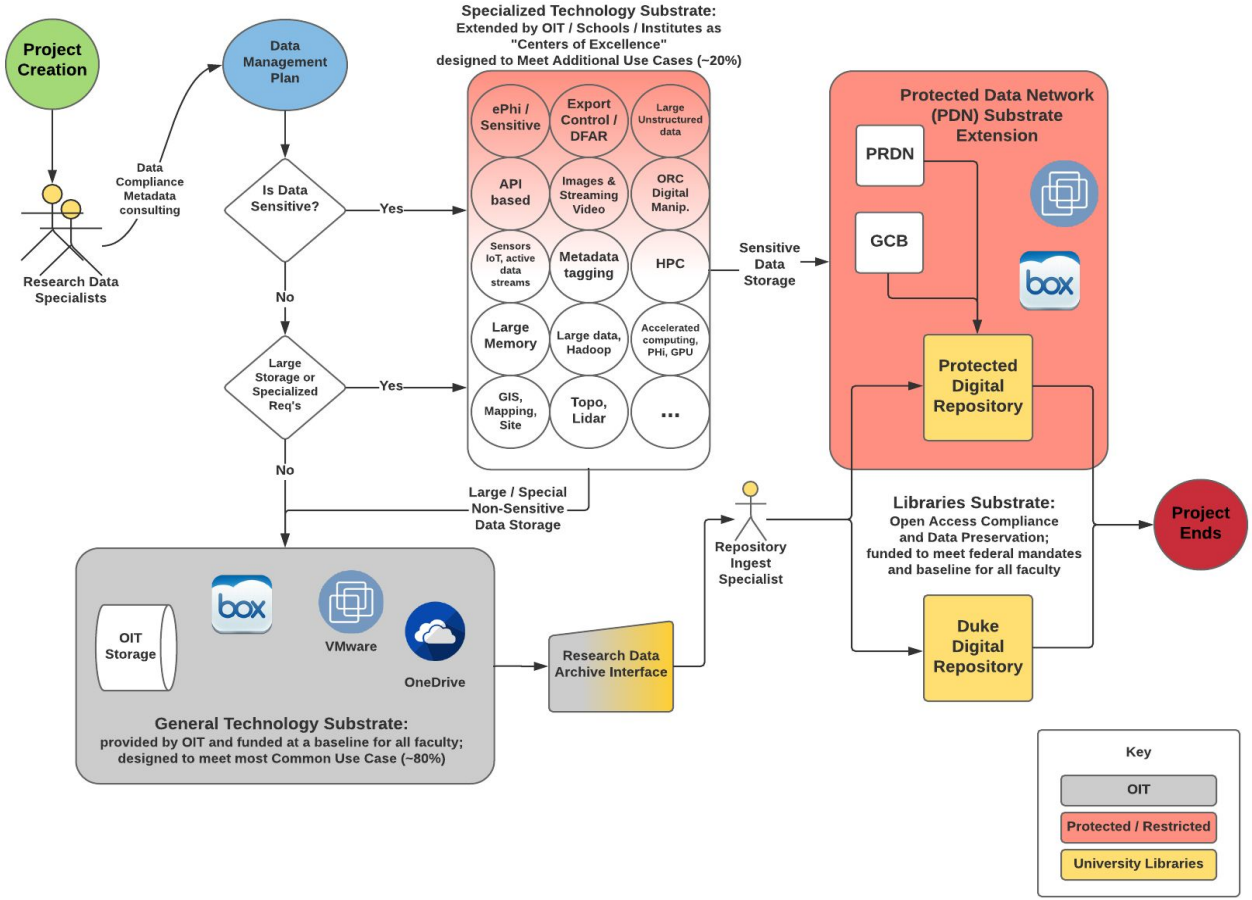
2015 - University Libraries, Trinity College of Arts & Sciences, and Office of Information Technology (OIT) each request significant funding increase for research data storage from Provost

2015 - Provost seeks white paper and charge for Faculty Working Group for Digital Research Data Services

2016 - Interdisciplinary faculty working group meets for 9 months and submits recommendation to Provost

DUKE DIGITAL RESEARCH DATA SERVICE MAP

Tracy Futhey, Tim McGeary | January 22, 2016




Education Program
Lifecycle Services
Curation Services

Education Program

Responsible Conduct of Research (RCR), students, staff, & faculty

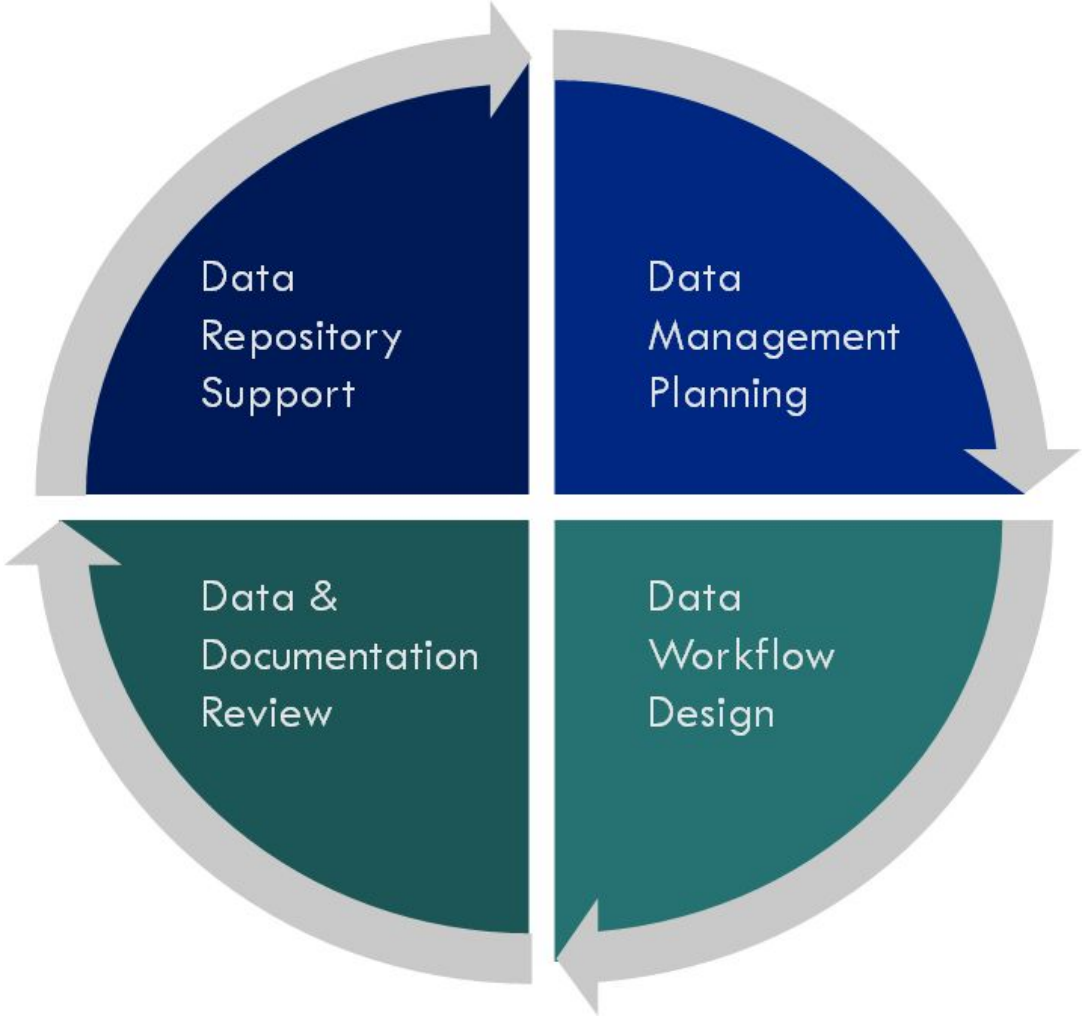
Tools, Best Practices, Reproducibility, Sensitive Data, Funder Compliance

12 workshops, 435 attendees
57% Sciences and Engineering, 20% Social Sciences, 6% Health Sciences, 6% Humanities

Office of Research Support, Research Computing, School of Medicine, etc.

All icons by VectorBakery from Noun Project.

Lifecycle Services



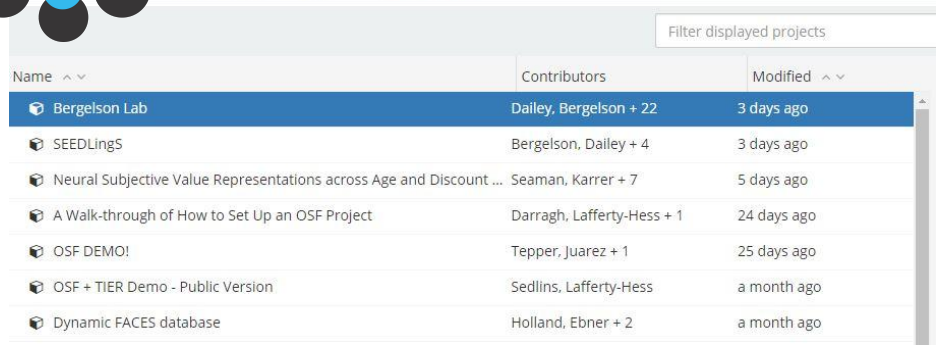
Lifecycle Services

The data will be deposited into the Duke Digital Repository (DDR) - repository.duke.edu - an openly accessible preservation archive maintained by the Duke University Libraries. The DDR will assign appropriate metadata (Dublin Core) for discoverability and provide a Digital Object Identifier (DOI) for persistent access and unique identification of the data.

Lifecycle Services

Duke UNIVERSITY

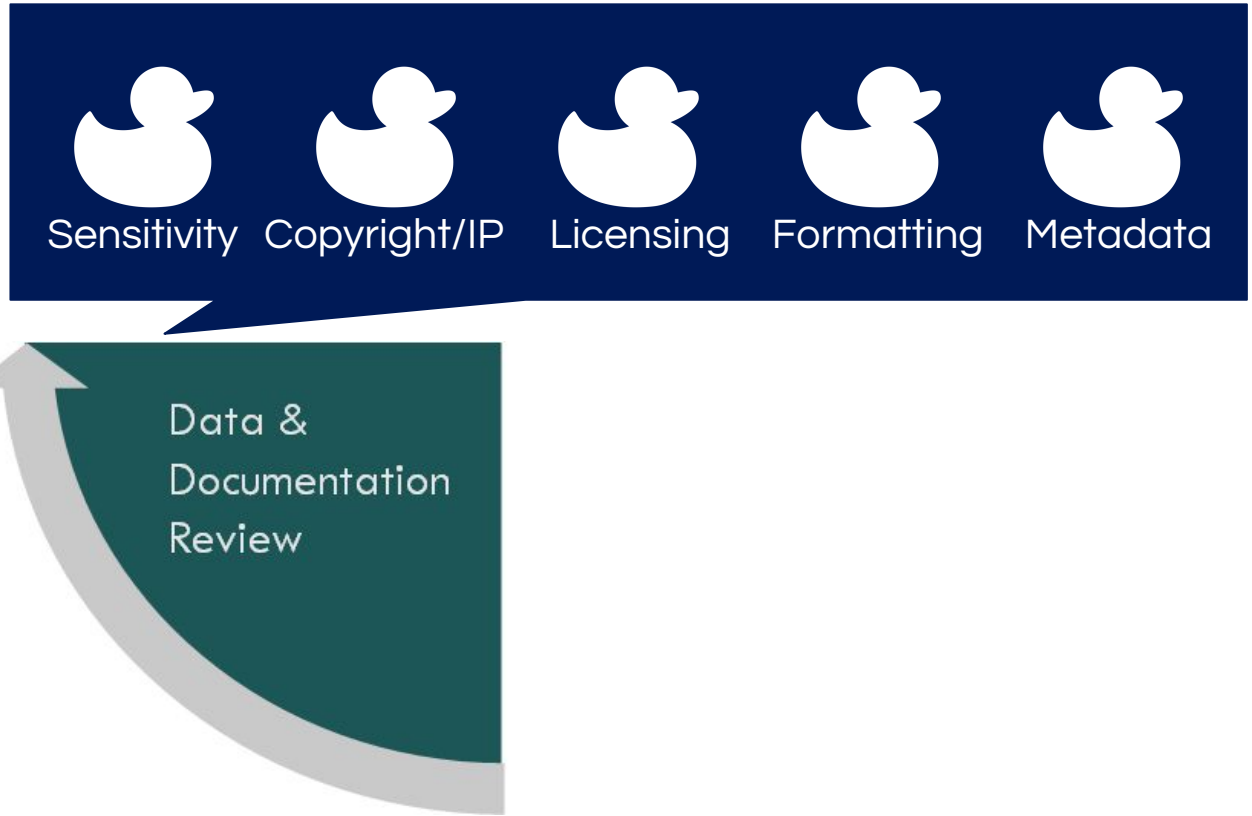
A research data service provided by Duke Libraries.



Data
Workflow
Design

Lifecycle Services

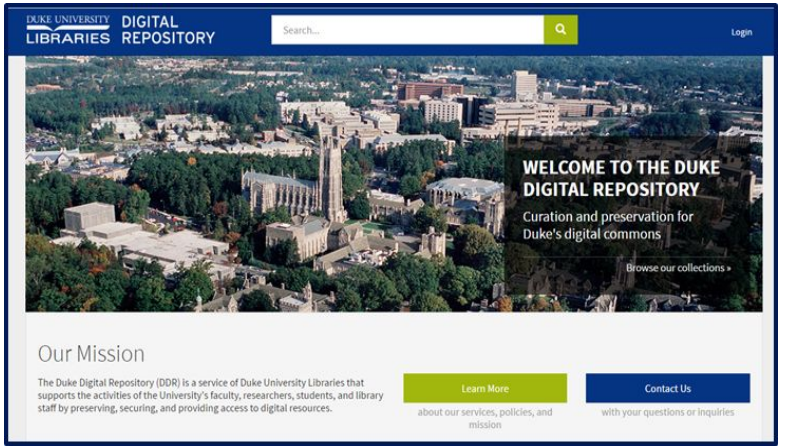
Icon by Christopher T. Howlett from Noun Project.

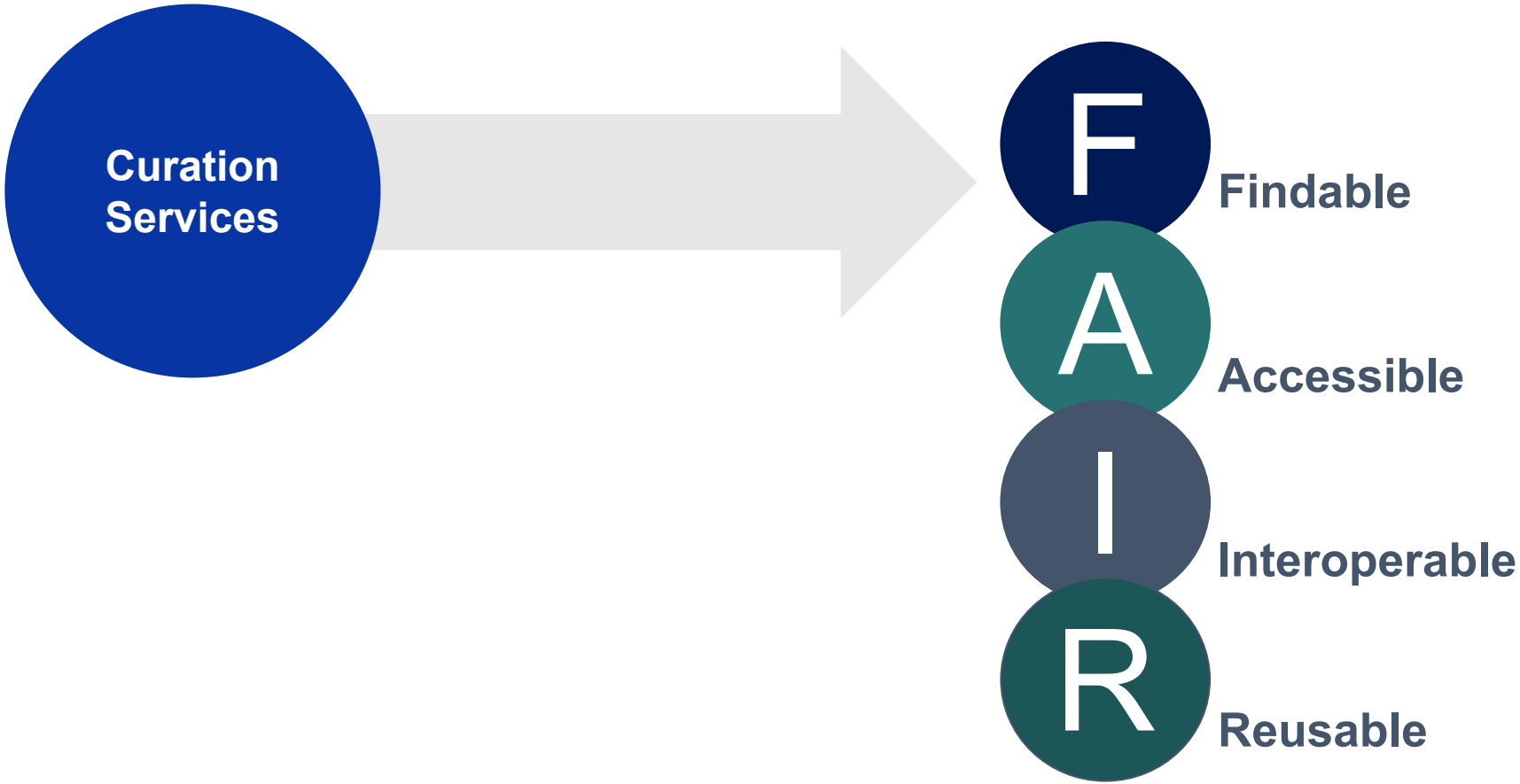


Lifecycle Services

Data
Repository
Support

https://library.duke.edu/ddr/research-data





Wilkinson, M. D.. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(160018). doi:10.1038/sdata.2016.18

Data Curation Pipeline @ Duke

Deposit

- ✓ Submit data, documentation, & metadata
- ✓ Select Creative Commons waiver/license

Data Curation Pipeline @ Duke

Deposit

- ✓ Submit data, documentation, & metadata
- ✓ Select Creative Commons waiver/license

Review

- ✓ Check documentation completeness
- ✓ Assess disclosure risks
- ✓ Assess formats

Data Curation Pipeline @ Duke

Deposit

- ✓ Submit data, documentation, & metadata
- ✓ Select Creative Commons waiver/license

Review

- ✓ Check documentation completeness
- ✓ Assess disclosure risks
- ✓ Assess formats

Ingest

- ✓ Transform files
- ✓ Arrange files
- ✓ Generate metadata
- ✓ Assign license

Data Curation Pipeline @ Duke

Deposit

- ✓ Submit data, documentation, & metadata
- ✓ Select Creative Commons waiver/license

Review

- ✓ Check documentation completeness
- ✓ Assess disclosure risks
- ✓ Assess formats

Ingest

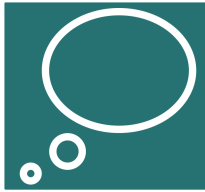
- ✓ Transform files
- ✓ Arrange files
- ✓ Generate metadata
- ✓ Assign license

Publish

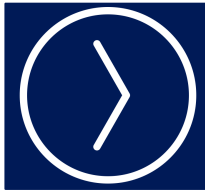
- ✓ Assign DOI
- ✓ Review & test
- ✓ Finalize administrative metadata

Curation Services

Disciplines served



Stories from our first year



Next steps

All icons by iconsmind.com from Noun Project.

Lessons Learned

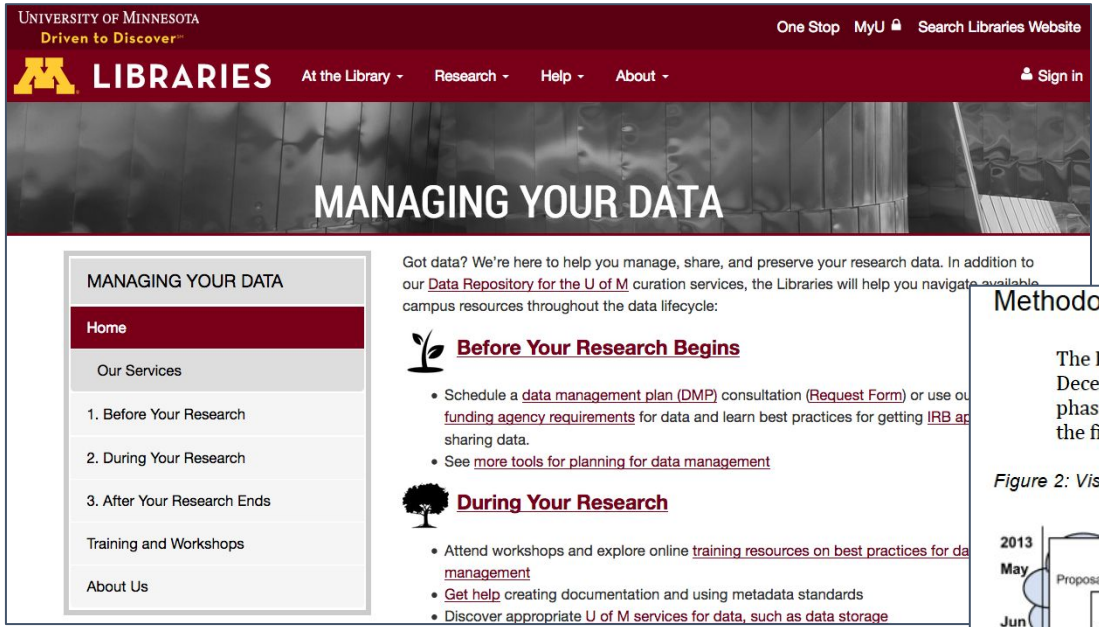
Value of looking towards peers for benchmarking and evaluation

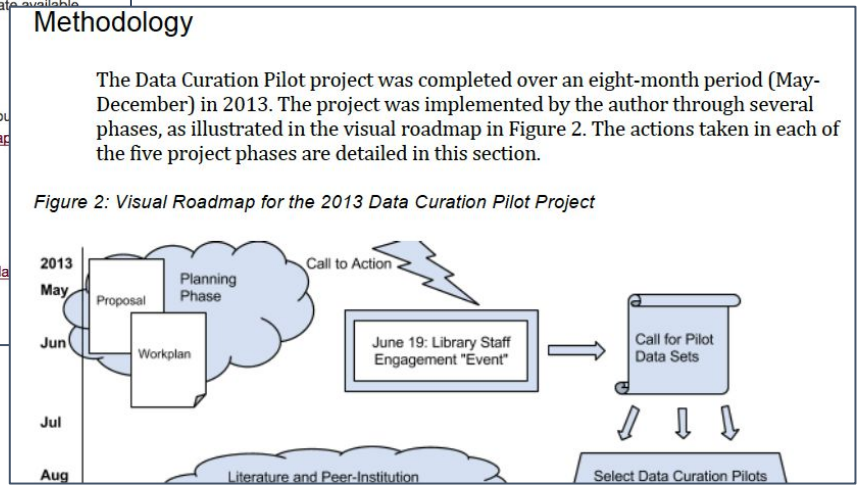
Further enhancing the FAIRness of data will benefit from in-depth subject and data type expertise

Collaborating both internally and externally facilitated program development

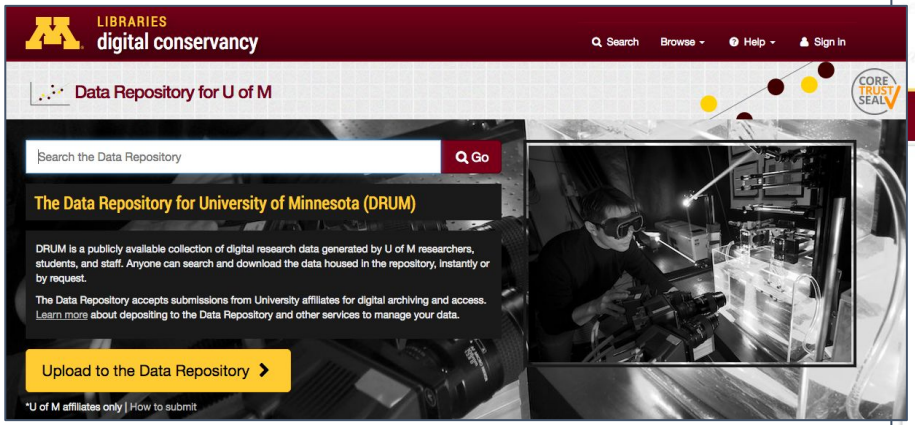
University of Minnesota / Data Curation Network

Origins, campus policy, and engagement





Origins, campus policy, and engagement



[UofM RDM policy](#) (January 2015)

Data Repository for the U of M (DRUM)

Data Repository for U of M

University Digital Conservancy Home / University of Minnesota / Data Repository for U of M (DRUM) / View Item

Cross-Cultural Interaction and Migration: Retheorizing Greek Colonization in the 7th -4th Centuries BC with the Exchange of Attic Figured Pottery

Faulkner-Gentry, Ivy (2018)

Published Date
2018-03-29

Author Contact
Faulkner-Gentry, Ivy (faul0078@umn.edu)

Type
Dataset
Observational Data
Spatial Data

Abstract
This data is part of a doctoral dissertation study on the exchange of Attic figured pottery between Greek settlers and local populations in Western Europe and the northern Black Sea in the 7th-4th centuries BC. This data compiles instances of Attic painted pottery in these regions with deposition context information as well as item descriptions and references. This data is also visualized geographically using ArcGIS Online (Esri), the results of which are shared via pdf and shapefiles. This data can be used to analyze inter- and intra-regional interaction and may form the foundation for expanded research on Greek migration in the Archaic period to other regions.

License
Attribution-NonCommercial-ShareAlike 3.0 United States

Data Repository for U of M

University Digital Conservancy Home / University of Minnesota / Data Repository for U of M (DRUM) / View Item

Predictors and benefits of microhabitat selection for offspring deposition in golden rocket frogs

Pettitt, Beth A; Bourne, Godfrey R; Bee, Mark A (2018)

Persistent link to this item
https://doi.org/10.13020/D6638B
http://hdl.handle.net/11299/195165

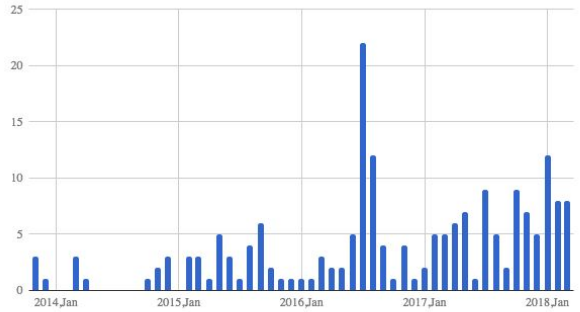
Services
Full Metadata (xml)
View Usage Statistics

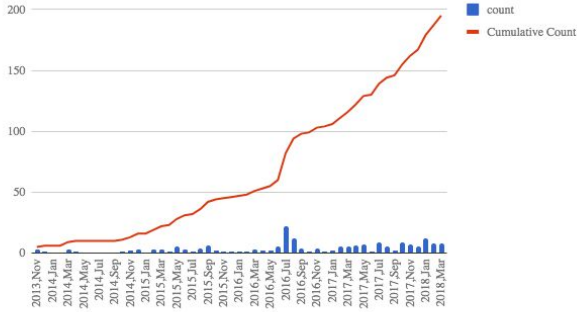
Published Date
2018-04-02

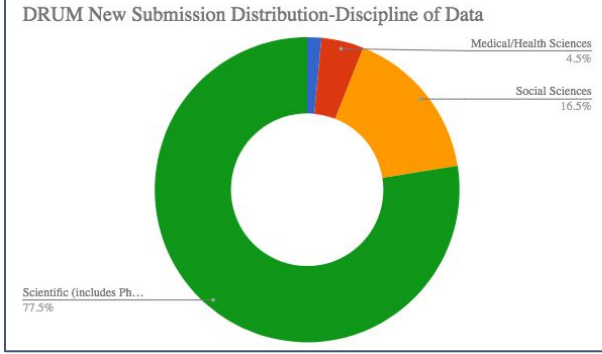
Author Contact
Bee, Mark A (mbee@umn.edu)

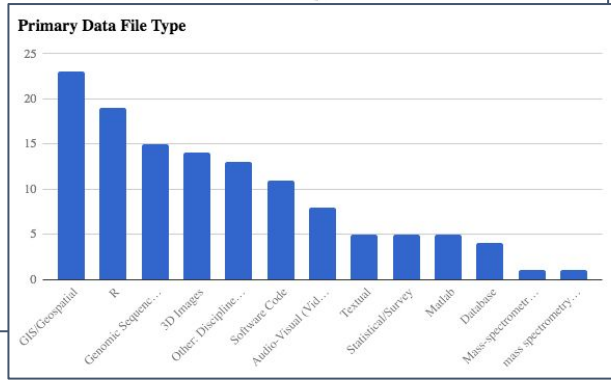
Type
Dataset
Field Study Data

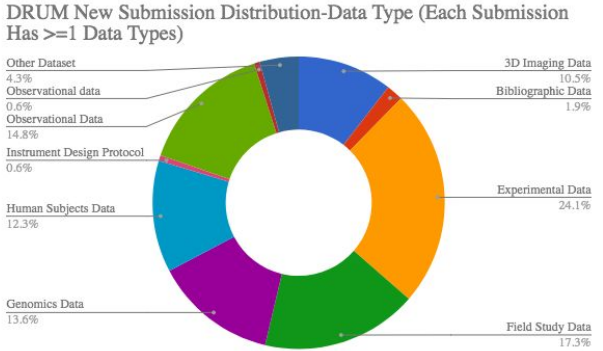
Abstract
This data set originates from a field study of habitat selection by parental frogs. In many tropical frogs, offspring development and survival potentially depend on microhabitat features associated with sites that parents select for oviposition and tadpole rearing. This study investigated the importance of microhabitat features in the selection of oviposition sites versus tadpole rearing sites, as well as in determining offspring survival, in the golden rocket frog, Anomaloglossus beebei. Endemic to Guyana, this species exhibits biparental care and exclusively uses phytotelmata in bromeliads for oviposition and tadpole rearing. The data included here were used in model based inference to evaluate evidence for the hypotheses that (1) parents prioritize different microhabitat features in selecting phytotelmata suitable for oviposition versus tadpole rearing and (2) microhabitat selection can adaptively promote offspring survival. The dataset includes descriptions of bromeliad size, phytotelm height, leaf angle (indicative of location within











Usage Statistics for Collection: Data Repository for U of M (DRUM)

Total Downloads (2014-Present)

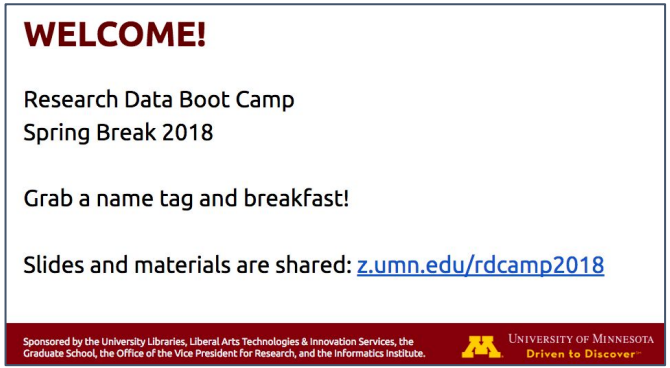
	File Downloads
Data Repository for U of M (DRUM)	85948

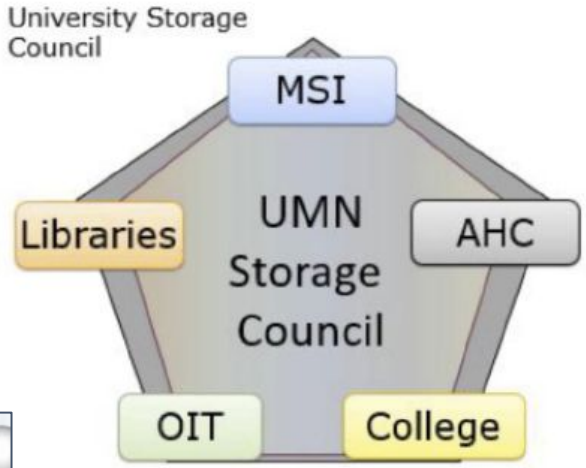
Data curation staff @ UMN

- Data Management and Curation initiative (now program) **lead**, 2013: Lisa Johnston
- **Humanities** data curator
  - 2013 CLIR two-year fellow digital humanities: Justin Schell
  - Continuous appointment librarian 2016: Director, Digital Arts Sciences and Humanities program (DASH), Ben Wiggins
- **Social sciences** data curator
  - 2015 CLIR two-year fellow social science data: Alice Motes
  - Continuous appointment librarian posting anticipated 2018
- **Health** data curator, new position 2015, Public Health Liaison and Data Curation Specialist: Shanda Hunt
- **Spatial data** analyst/curator, jointly funded with campus U-Spatial project, Melinda Kernik
- Digital Repositories **Archivist**, new position 2016: Valerie Collins
- **Scientific** data curator
  - Graduate student 50% time 2015-2016
  - Continuous appointment librarian 2017: Biosciences Librarian and Scientific Data Specialist, Katie Wilson

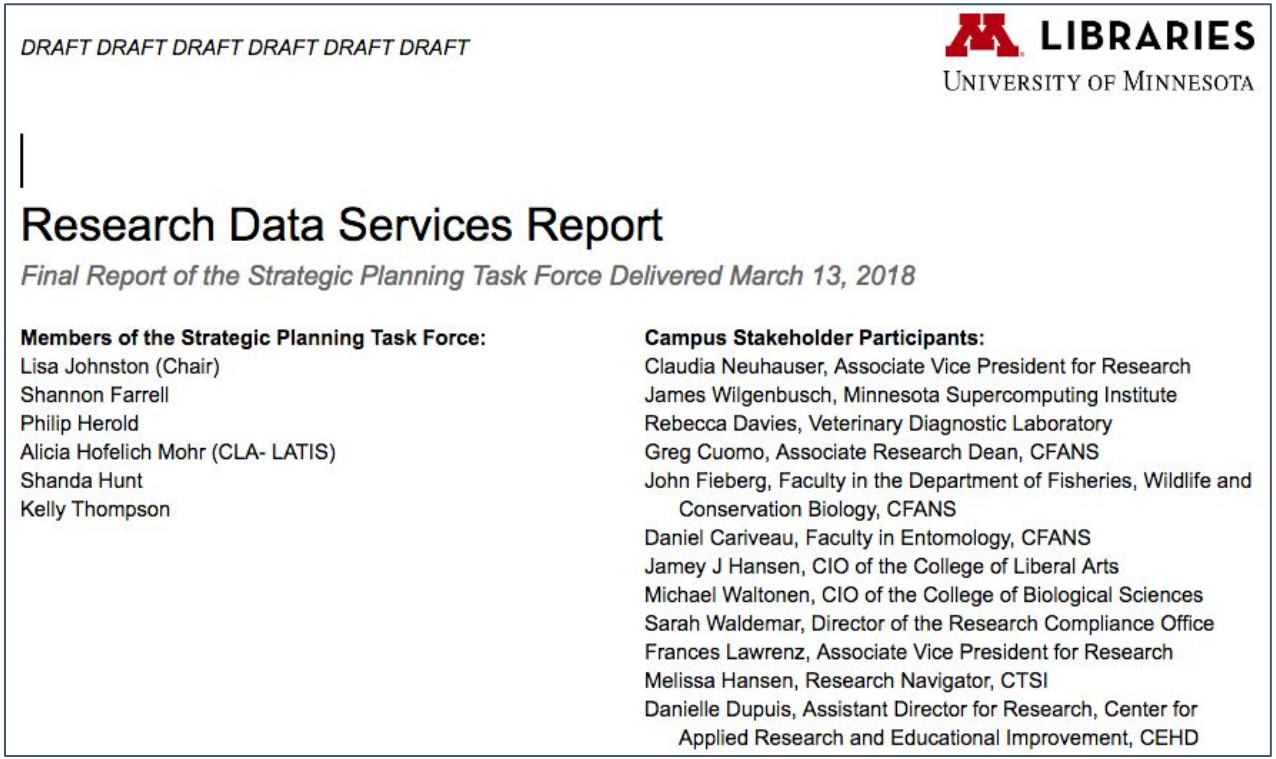
Also critical: repository developers/technologists, digital preservation librarian, University Archivist, Research Services Coordinators and liaisons, colleagues in collegiate technology units

Campus uptake, engagement, evolution





New strategic plan for RDS in draft



DRAFT DRAFT DRAFT DRAFT DRAFT DRAFT

LIBRARIES
UNIVERSITY OF MINNESOTA

Research Data Services Report

Final Report of the Strategic Planning Task Force Delivered March 13, 2018

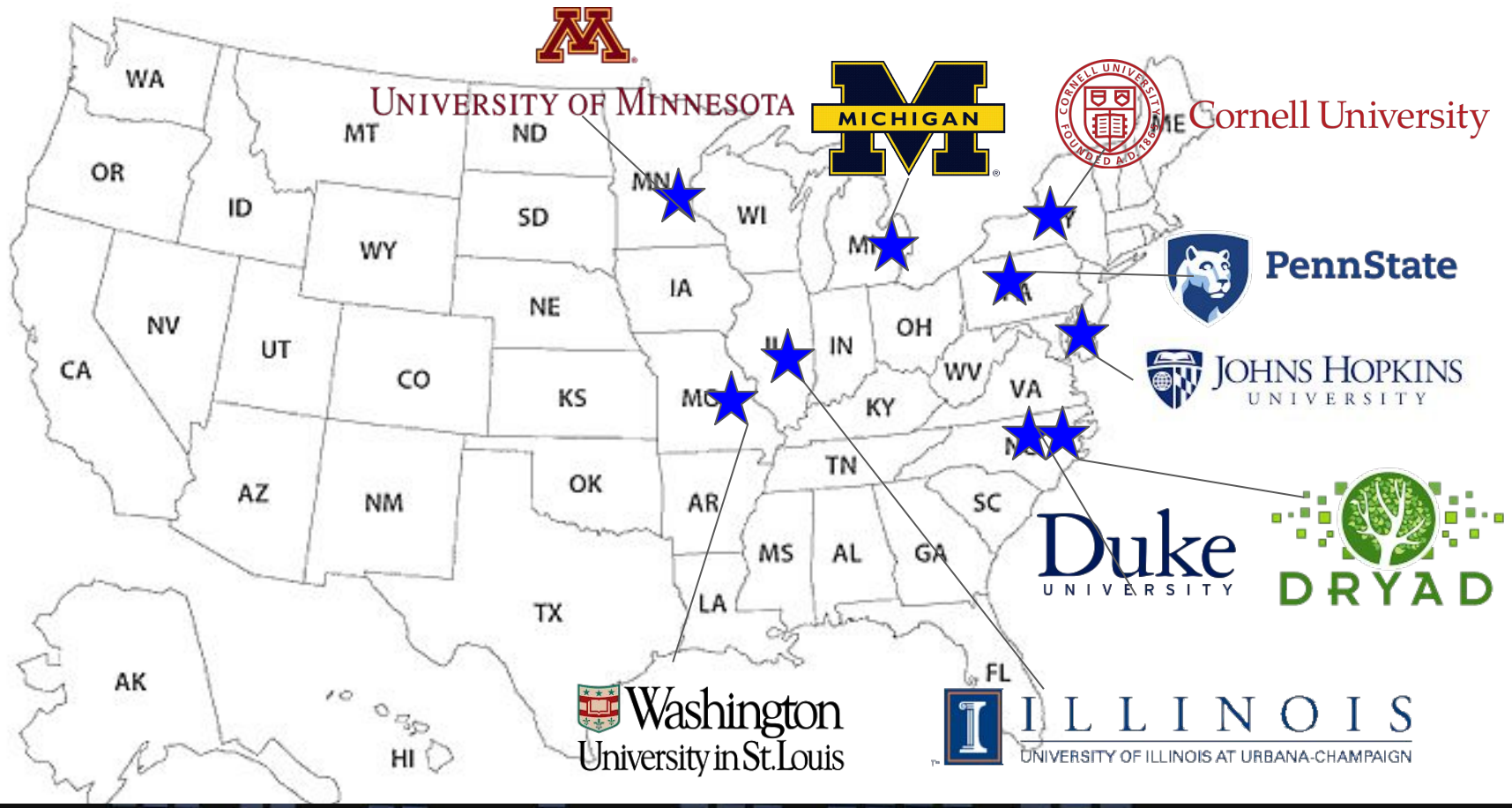
Members of the Strategic Planning Task Force:
Lisa Johnston (Chair)
Shannon Farrell
Philip Herold
Alicia Hofelich Mohr (CLA- LATIS)
Shanda Hunt
Kelly Thompson

Campus Stakeholder Participants:
Claudia Neuhauser, Associate Vice President for Research
James Wilgenbusch, Minnesota Supercomputing Institute
Rebecca Davies, Veterinary Diagnostic Laboratory
Greg Cuomo, Associate Research Dean, CFANS
John Fieberg, Faculty in the Department of Fisheries, Wildlife and Conservation Biology, CFANS
Daniel Cariveau, Faculty in Entomology, CFANS
Jamey J Hansen, CIO of the College of Liberal Arts
Michael Waltonen, CIO of the College of Biological Sciences
Sarah Waldemar, Director of the Research Compliance Office
Frances Lawrenz, Associate Vice President for Research
Melissa Hansen, Research Navigator, CTSI
Danielle Dupuis, Assistant Director for Research, Center for Applied Research and Educational Improvement, CEHD

Alfred P. Sloan
FOUNDATION

The Data Curation Network

Lisa Johnston *University of Minnesota*
Jake Carlson *University of Michigan*
Cynthia Hudson--Vitale *Washington Univ*
Heidi Imker *University of Illinois*
Wendy Kozlowski *Cornell University*
Robert Olendorf *Penn State University*
Claire Stewart *University of Minnesota*
Mara Blake *Johns Hopkins University*
Joel Herndon *Duke University*
Elizabeth Hull *Dryad Data Repository*
Timothy M. McGeary *Duke University*



Data Curation Network datacurationnetwork.org

Well curated data are more valuable.

The skills and expertise required to curate data cannot be fully automated nor reasonably be provided by a few experts siloed at single institutions.

The Data Curation Network (DCN) addresses this challenge by collaboratively sharing data curation staff across a network of partner institutions and data repositories.

Steps in building the Network

1. **DCN planning phase research (2016-2017)**

2. **Implementation launch in Spring 2018**
   a. DCN staffing model + Advisory panel
   b. DCN training/networking events
   c. DCN workflow and C-U-R-A-T-E steps
   d. Assessment Plan

3. **Grow the DCN beyond our grant-funded phase to a sustainable entity**
   a. Criteria for new partners
   b. Proposed financial model (alliance curation-as-service)

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

1. **Compared local policy,** technologies, and workflows across the 6 planning phase institutions;
2. **Held six focus groups** with researchers on what data curation activities were important;
3. **Ran controlled pilots** of data curation workflows with 17 data curators to ID issues;
4. **Surveyed the 124** ARL institutions to gauge support for data curation services;
5. **Researched cost recovery** models for sustainable data curation and repository services;
6. **Held information exchanges** with leaders of successful collaboration projects;
7. **Analyzed one-year** of data types, disciplines, frequency, and curation levels needed vs taken).

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Baseline Assessment

	Pre-ingest Curation?		Mediated vs Self-deposit?		Post-ingest curation			
Workflow Steps by Institution	Consult only	Staging Area for deposit	Mediated deposit	Self-deposit	As needed	Review metadata only	Review files and metadata	Add DOI
Minnesota	X			X			X	X
Cornell	X		X*	X			X	X*
Illinois	X			X			X*	X
Michigan	X			X			X*	X*
Penn State	X			X				
Wash U	X		X	X			X	X

"Data Curation Network: How Do We Compare? A Snapshot of Six Academic Library Institutions' Data Repository and Curation Services." Journal of eScience Librarianship 6(1): e1102. https://doi.org/10.7191/jeslib.2017.1102.

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

How would we deal with conflicting policy issues?

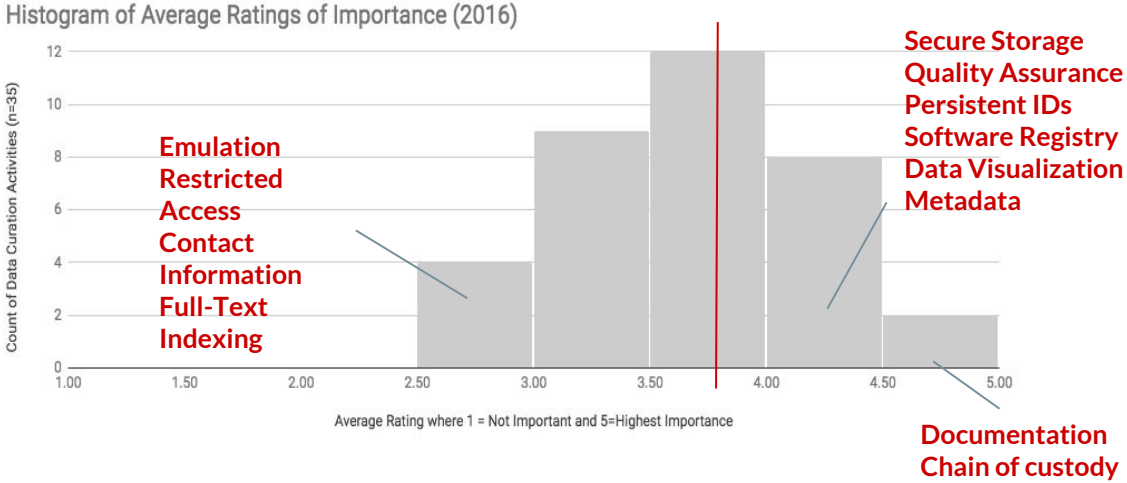
What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

Researcher Focus Groups (n=91)

Ave Rating = 3.7 out of 5



(in press) "How Important Are Data Curation Activities to Researchers? Gaps and Opportunities for Academic Libraries," *Journal of Librarianship and Scholarly Communication.*

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

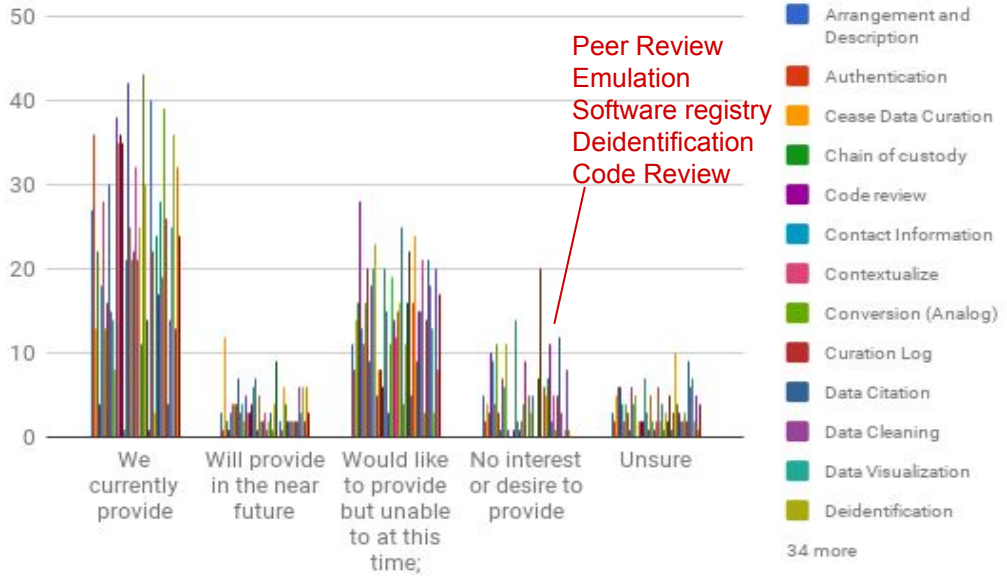
How would we deal with conflicting policy issues?

What do researchers actually need our help with? Will they care if curation is distributed?

Can I trust someone else to curate our data? What about quality control?

What skills do we need? What types of data sets are deposited into our data repositories? How long does curation take?

ARL Institutions Survey (n=80)



SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017.
http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

Planning Phase (2016-2017)

Alfred P. Sloan FOUNDATION

What are the measure of success?

How can we grow and sustain the Network beyond the grant-funding period?

Information Exchanges

2CUL

DURASPACE

portage
SERVICES PARTAGÉS POUR LES DONNÉES DE RECHERCHE
SHARED STEWARDSHIP OF RESEARCH DATA

SEAD

University of California CDL California Digital Library

TDL.ORG TEXAS DIGITAL LIBRARY

ISB International Society for Biocuration

CURATING for REPRODUCIBILITY

FDA U.S Food and Drug Administration

DCN Implementation

9 Institutions
- 8 Academic Libraries
- 1 General Data Repository

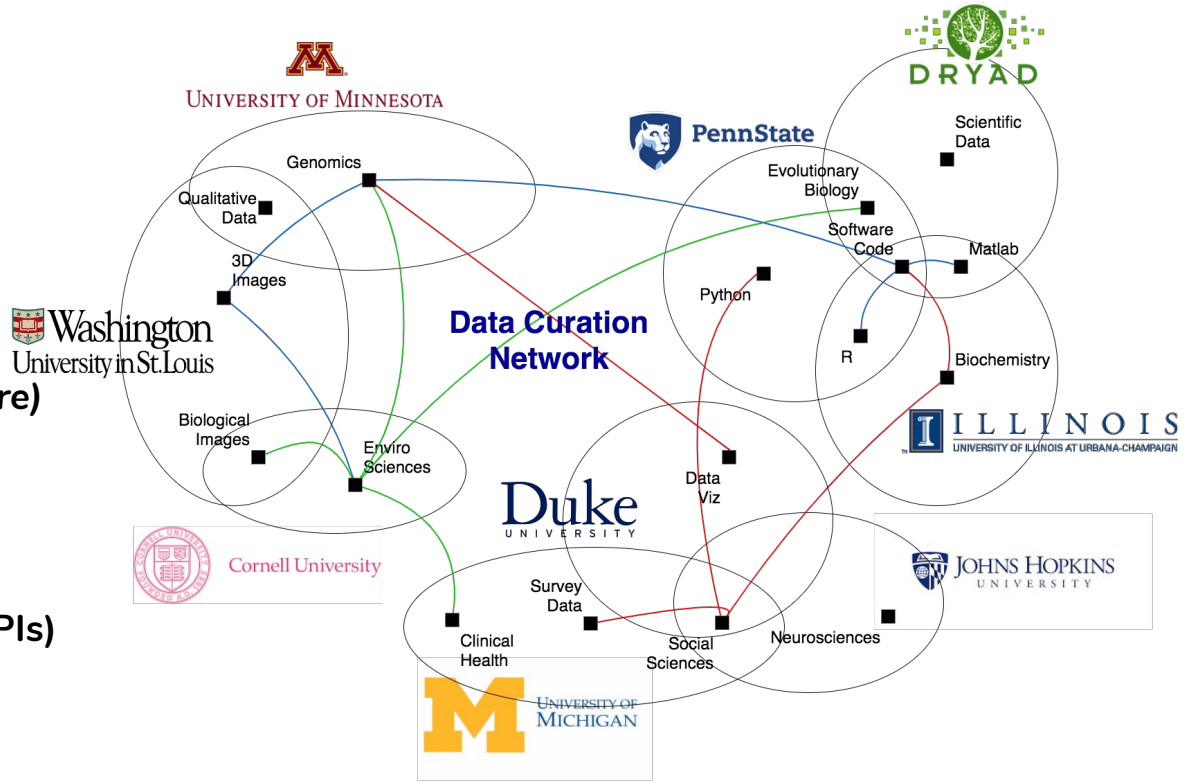
19 Data Curators

1 Project Coordinator ***(new hire)***

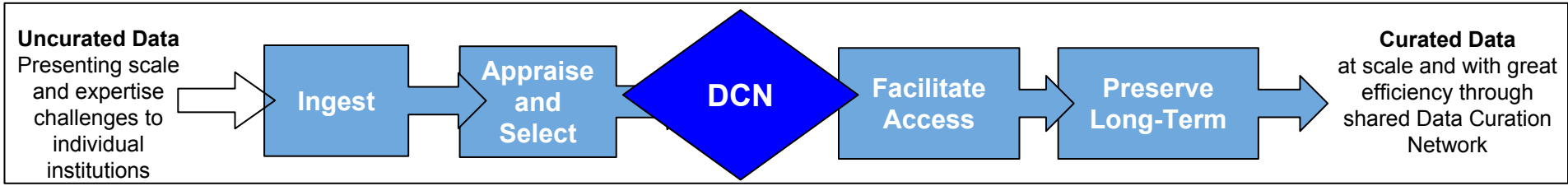
1 Program Director **(PI)**

8 DCN Representatives **(CO-PIs)**

2 Admin Leads

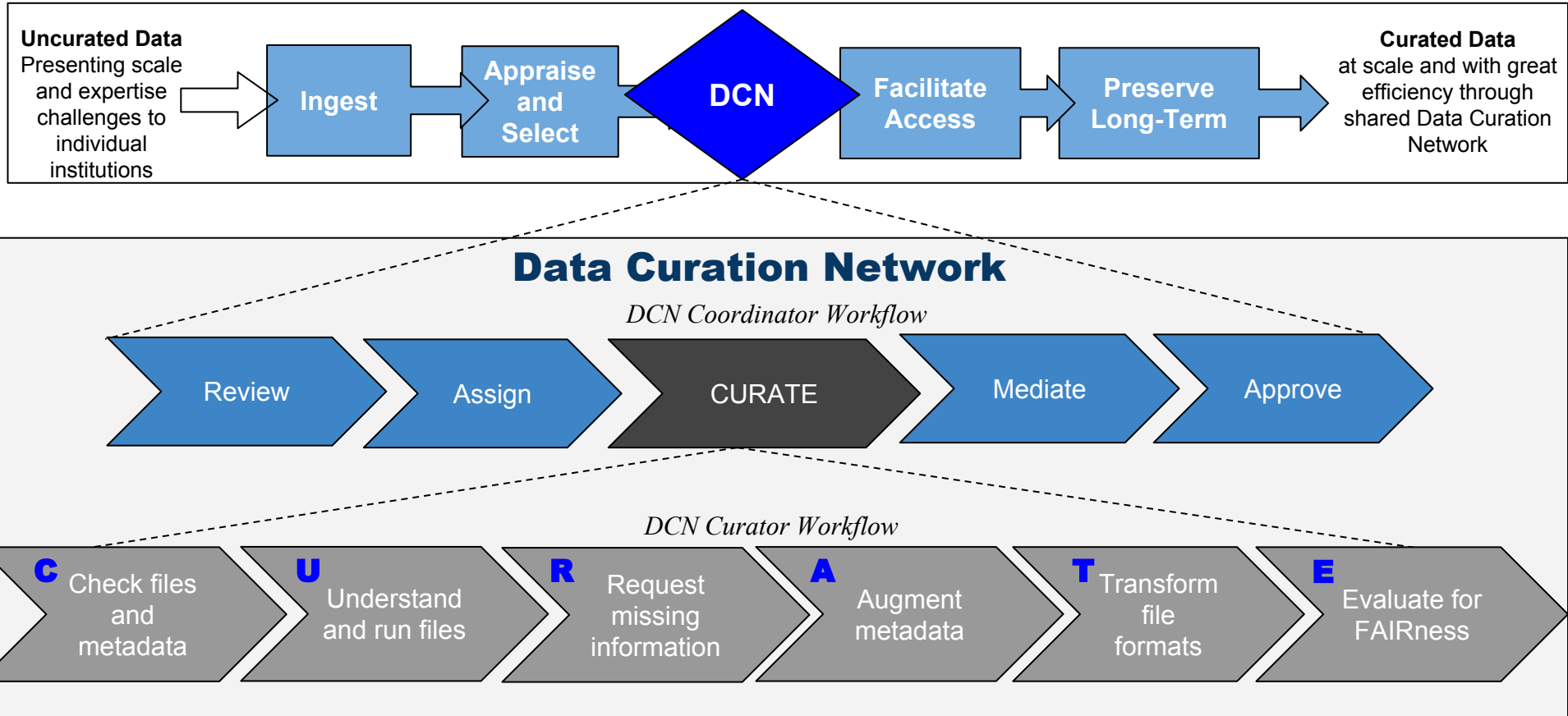


DCN Workflow



- Researchers deposit like normal
- DCN functions as a microservice layer (the “human layer in your repository stack”)
- Local institution maintain full responsibility for all technical functionality (eg. storage) and authority for local decision-making (what to ingest, how long to retain, etc.)
- Seamlessly integrates into all repository systems (Samvera, Fedora, DSpace, etc.)

DCN Workflow



CURATE Steps in DCN Workflow

DCN Curators will take **CURATE** steps for each data set, that includes:

- **C** **Check** data files and read documentation
- **U** **Understand** the data (try to), if not…
- **R** **Request** missing information or changes
- **A** **Augment** the submission with metadata for findability
- **T** **Transform** file formats for reuse and long-term preservation
- **E** **Evaluate** and rate the overall submission for FAIRness.

DCN Implementation (2018-2020)

Assessment Plan (two-prong)

Is a networked approach to curating research data more efficient?

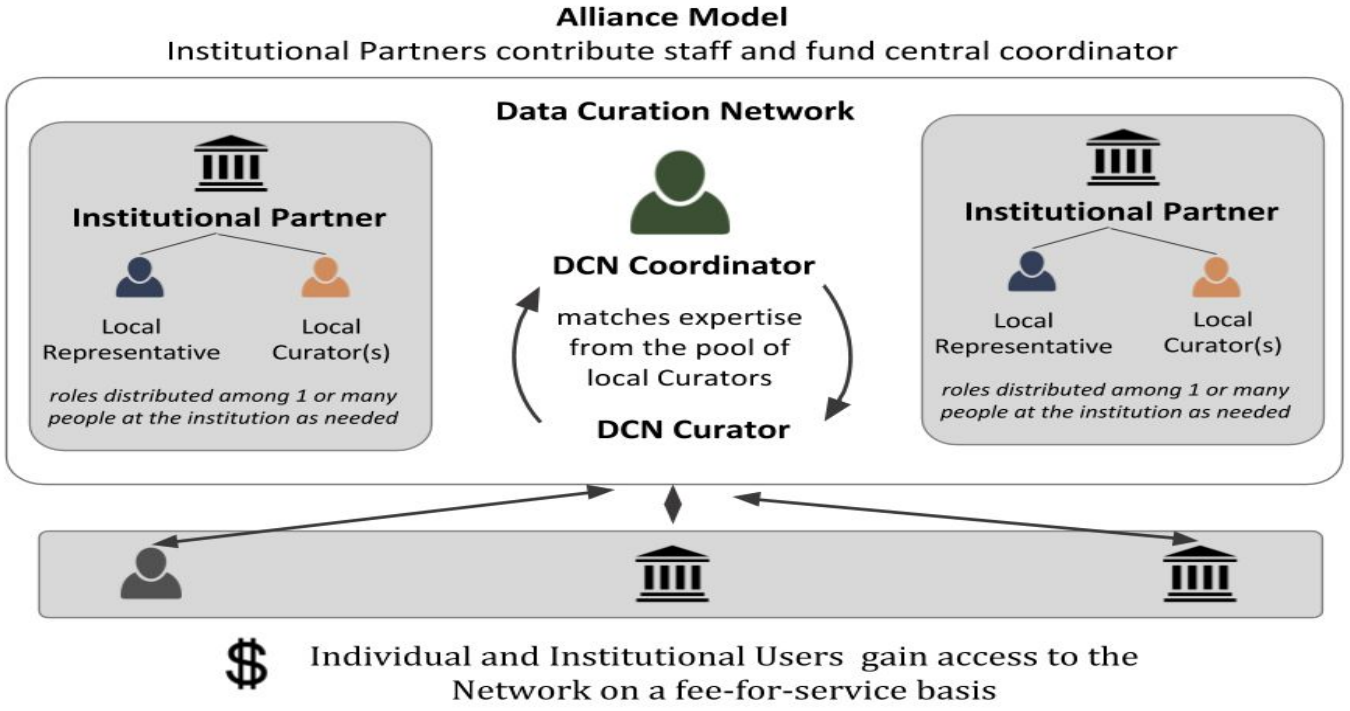
- Number of datasets
- Frequency (high-volume time periods, etc.)
- Variety (data file formats; range of disciplines)
- Efficiency (time, costs)

Are curated data are more valuable?

- Track reuse indicators (download counts, citations, alt-metrics)
- Implement a DCN registry
- Apply badges and metadata to signal that data sets curated by the DCN are FAIR.

*In Year 3, the DCN will begin transitioning to a **self-sustaining service model** where institutional and disciplinary partners contribute data curation staff and central operations costs are offset by users of the Network.*

Data Curation-as-service



Value proposition

Stakeholder	*Benefits*
Academic libraries with existing data curation services	Gain access to data curation expertise in more disciplines/formats than locally available
Academic libraries with limited to no resources for data curation services	Are able to provide critical new data curation services when local resources are limited (without needing to hire);
Disciplinary- and general-subject data repositories	Receive better, more valuable data submissions from DCN partner institutions and customers; Have potential to partner with the DCN to expand the scope of curation support for new and/or less frequently encountered data types

DCN FAQ

- Do researchers actually value these services?
- Won't researchers curate their own data?
- Is it another community of practice?
- Aren't you all large research libraries? How about other kinds of libraries?
- Can't some of this (data curation) be automated?
- Why aren't you also sharing a repository?
- Why charge $ for this?

What other questions do you have?

Thank you!

Tim McGeary
tim.mcgeary@duke.edu

Claire Stewart
cstewart@umn.edu

Sophia Lafferty-Hess
sophia.lafferty.hess@duke.edu

Jennifer Darragh
jennifer.darragh@duke.edu